Interpreting Subgroup Analyses in Clinical Trials

Allan Donner, Phd, FRSC
Professor, Department of Epidemiology and Biostatistics
Director, Biometrics
Robarts Clinical Trials
Robarts Research Institute
University of Western Ontario
London, Canada
## Comparison of Regimens for treating HIV infected Patients

Percent of Patients with HIV RNA levels <50 c/ml

<table>
<thead>
<tr>
<th></th>
<th>Raltegravir (n=263)</th>
<th>Efavirenz (n=258)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>92%</td>
<td>89%</td>
<td>.33</td>
</tr>
<tr>
<td>Latin America</td>
<td>91%</td>
<td>84%</td>
<td></td>
</tr>
<tr>
<td>Southeast Asia</td>
<td>94%</td>
<td>93%</td>
<td></td>
</tr>
<tr>
<td>North America</td>
<td>90%</td>
<td>92%</td>
<td></td>
</tr>
<tr>
<td>Europe/Australia</td>
<td>94%</td>
<td>91%</td>
<td></td>
</tr>
</tbody>
</table>

Lennox et al (2009)
Comparison of regimens for treating patients with acute MI

MORTALITY WITHIN 90 DAYS OF PATIENT ENTRY

<table>
<thead>
<tr>
<th>Group</th>
<th>Placebo (n=697)</th>
<th>Metoprolol (n=698)</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Patients</td>
<td>8.9</td>
<td>5.7</td>
<td>.03</td>
</tr>
<tr>
<td>Age 40-69</td>
<td>8.1</td>
<td>5.1</td>
<td>.04</td>
</tr>
<tr>
<td>Age 70-74</td>
<td>15.7</td>
<td>11.6</td>
<td>&gt;.20</td>
</tr>
<tr>
<td>Age 40-64</td>
<td>5.7</td>
<td>4.5</td>
<td>&gt;.20</td>
</tr>
<tr>
<td>Age 65-74</td>
<td>14.8</td>
<td>8.1</td>
<td>.03</td>
</tr>
<tr>
<td>Definite MI</td>
<td>13.7</td>
<td>9.0</td>
<td>.05</td>
</tr>
<tr>
<td>No Definite MI</td>
<td>2.1</td>
<td>1.3</td>
<td>&gt;.20</td>
</tr>
</tbody>
</table>

Hjalmarson et al.(1981)
INTERPRETATIONAL PROBLEMS

Type I error rates:

Since the number of subgroups of interest may be large, the overall probability of detecting at least one false significant difference may be far greater than nominal.
Suppose separate significance tests are performed within each of \( G \) patients subgroups. If each test is performed at the 5\% level of significance \( \alpha \) and the treatments are identical then the probability of obtaining at least one significant result is:

\[
1 - (1 - \alpha)^G
\]

e.g., if \( \alpha = .05 \) and \( G=10 \), then this probability is 0.40.
Type II error rates:

Since the sample sizes for many subgroup comparisons tend to be small, the probability of failing to detect a substantively important subgroup effect may be high.

Thus suppose the overall treatment effect is significant while the corresponding effect in a selected subgroup of patients is non-significant.

It is misleading to conclude that the results for this subgroup differ from the overall results.
Estimation Bias

Treatment effects that are reported BECAUSE they are statistically significant (i.e. because they are extreme) will on the average be less extreme when the treatment is applied to a new similar set of subjects.
**Example 1:** Effect of Nimodipine Treatment on Neurological Outcome After Ischemic Stroke: Subgroup Analyses by Delay to Start of Treatment

<table>
<thead>
<tr>
<th>Patient group</th>
<th>No. of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Trt</td>
</tr>
<tr>
<td>All</td>
<td>1811</td>
</tr>
<tr>
<td>Start ≤ 12 hrs</td>
<td>330</td>
</tr>
<tr>
<td>Start 13-24 hrs</td>
<td>451</td>
</tr>
<tr>
<td>Start &gt; 24 hrs</td>
<td>803</td>
</tr>
</tbody>
</table>

Odds ratios for unfavorable outcome with 95% confidence interval

Nimodipine better
Nimodipine worse
“Those starting within 12h showed the strongest beneficial effect. Those starting after 12h showed a slightly negative effect”.

J.P. Mohr et al. (1994)
"Is it biologically plausible that treatment within 12 hours is beneficial but later treatment is harmful? Such qualitative interactions are rare in medicine.

The difference between the effects of early and late treatment may similarly have been due to chance effects. Where the subgroup analyses have not been predefined, even greater caution is needed in interpreting the results".

Counsell et al. (1994)
## Example 2:

Comparison of Combination Therapy (Zalcitabine and Zidovudine) vs. Zidovudine on Patients with HIV Disease

<table>
<thead>
<tr>
<th>Pretreatment CD4 Cell Count</th>
<th>Favor Combination</th>
<th>Favor Zidovudine</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 – 150</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 50</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

95% confidence interval for relative risk

These results indicate the treatment comparison is significant for patients with a CD4 cell count exceeding 150 but not significant for patients with a CD4 cell count of 150 or less.

Fischl et al. (1995)
"The investigators were criticized for presenting results that focused on subgroup analysis. The criticism was twofold: first for promoting the subgroup analysis as a positive result when the primary analysis was nonsignificant and second for focusing on this subgroup analysis that was not described in the original protocol, but was added to the protocol partway through the study".

Korzun and Chaloner (1995)
Example 3:

Do trials sponsored by the pharmaceutical industry show more favorable results?

<table>
<thead>
<tr>
<th>Trials sponsored by Industry</th>
<th>No. of Trials (No. of Patients)</th>
<th>Relative Odds of Relapse on Clozapine</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trials sponsored by Industry</td>
<td>13 (980)</td>
<td>0.50</td>
<td>(0.3 – 0.7)</td>
</tr>
<tr>
<td>Trials Not sponsored by Industry</td>
<td>10 (783)</td>
<td>0.40</td>
<td>(0.1 – 1.4)</td>
</tr>
</tbody>
</table>

“Our finding supports the concern that drug company involvement in clinical trials affects the outcome”

Wahlbeck and Adams (1999)
CRITIQUE

“An unequivocal difference in the sponsored trials and an equivocal difference in the unsponsored trials does not establish a difference between the two types of study. To claim that it does amounts to comparing subgroups on the basis of their p-values, and this is known to be flawed”.

Matthews (1999)
TEST OF INTERACTION

The purpose of tests of interaction is to demonstrate that the effect of treatment significantly varies across subgroups.

This approach is to be preferred to performing separate significance-tests within subgroups.
Randomized trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction: Cardiac mortality by sex

<table>
<thead>
<tr>
<th>Cardiac Mortality</th>
<th>Men</th>
<th></th>
<th>Women</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Intervention (n=458)</td>
<td>Control (n=445)</td>
<td>Intervention (n=234)</td>
<td>Control (n=239)</td>
</tr>
<tr>
<td>Died</td>
<td>11 (2.4%)</td>
<td>11 (2.5%)</td>
<td>22 (9.4%)</td>
<td>12 (5.0%)</td>
</tr>
<tr>
<td>Survived</td>
<td>447</td>
<td>434</td>
<td>212</td>
<td>227</td>
</tr>
<tr>
<td>Total</td>
<td>458</td>
<td>445</td>
<td>234</td>
<td>239</td>
</tr>
</tbody>
</table>

“The poor overall outcome for women, and the possible harmful impact of the intervention on women, underline the need for further research”.

Frasure-Smith et al. (1996)
Does the effect of intervention depend on sex?

**Test of Interaction:**

\[
\begin{align*}
\text{OR}_1 &= \frac{(11)(434)}{(11)(447)} = 0.97 \\
\text{OR}_2 &= \frac{(22)(227)}{(12)(212)} = 1.96 \\
\log_e(\text{OR}_1) &= 0.029514 \\
\log_e(\text{OR}_2) &= 0.67445
\end{align*}
\]
$\hat{V}\left(\log_e[OR_1]\right) = 0.18636$

$\hat{V}\left(\log_2[OR_2]\right) = 0.13791$

$$Z = \frac{-0.029514 - 67445}{\sqrt{0.18636 + 0.13791}} = \frac{-0.70396}{0.56945} = -1.24 (p = .21)$$

**CONCLUSION:** Insufficient evidence exists that effect of intervention depends on sex.
QUANTITATIVE VS. QUALITATIVE INTERACTION

Quantitative Interaction:

The magnitude, but not the direction, of the true treatment effect varies across subgroups.

Qualitative Interaction:

The direction of the true treatment effect varies across subgroups. This implies that one treatment is superior for some subsets of patients and the alternative treatment is superior for others.
Bonferroni procedure:

To declare a given subgroup comparison to be statistically significant at level $\alpha$, require $P < \alpha / G$, where $G$ is the total number of subgroup comparisons performed.

E.g., if $G = 10$, then to declare significance at $\alpha = .05$, require $P < .05/10 = .0050$.

This procedure is known to be very conservative, especially if the subgroups are not independent.
Summary

**Bonferroni Procedures**

- No distributional assumptions
- Require only P-values to implement
Tests of Interaction

- Provide more accurate approximations to true P-value.
- Can be extended through statistical modelling to adjust for covariates.
- Are associated in a natural way with estimates of effect.
A contemporary difficulty created by subgroup analyses: the lay press.

“Why not allow companies to cull the relevant data from existing studies when a certain subgroup is clearly of help?”

Editorial. Wall Street Journal
November 26, 2002
Questions to ask:

(i) Were the subgroup hypotheses formulated in advance?

(ii) Are the results biologically credible?

(iii) Are the results consistent with those from previous trials?

(iv) Were tests of interaction or other methods of controlling type 1 error rate performed?
### Subgroup analysis by astrological sign

<table>
<thead>
<tr>
<th>Birth Sign</th>
<th>Percent Reduction in Odds of Death</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scorpio</td>
<td>48%</td>
<td>&lt; .04</td>
</tr>
<tr>
<td>All others</td>
<td>12%</td>
<td>Not Significant</td>
</tr>
</tbody>
</table>

Collins, R. et al. (1987)
INTERNAL CROSS-VALIDATION (SAMPLE-SPLITTING)

This technique allows subgroup hypotheses to be generated and tested in the same clinical trial. A portion of the data may be used to formulate hypotheses, with the remaining portion reserved for confirmatory testing.
**Advantage:** Has intuitive appeal and allows one to use judgement and prior information in selecting subgroups of interest.

**Disadvantage:** Not very efficient from a statistical point of view.
BAYESIAN METHODS

These methods, based on specifying prior probabilities of clinically important interactions, tend to shrink the point estimate of a subgroup effect toward the overall estimate of treatment effect.
**Advantage:** Prior beliefs concerning the presence of interactions are often based on solid evidence, and Bayesian methods allow these beliefs to formally influence clinical decision-making.

**Disadvantage:** There will inevitably be disagreements on the appropriate choice of prior probabilities, and thus with respect to the conclusions drawn. These methods, as well, are still in the developmental stage.
Hierarchy of credibility

(i) Subgroups specified in advance in protocol.
(ii) Subgroups implied by stratification factors.
(iii) Subgroups identified by other similar trials.
(iv) Subgroups identified through trial monitoring and tested in later subjects.
(v) Subgroups suggested by the data.
(vi) Subgroups categorized by outcome variables, i.e. events occurring after baseline.

Friedman, Furberg, DeMets
Fundamentals of Clinical Trials (1996)
### “Outcome by Outcome” analyses

<table>
<thead>
<tr>
<th>Dose Level</th>
<th>Total Group Number of Patients</th>
<th>Survival Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>78</td>
<td>77</td>
</tr>
<tr>
<td>Moderate</td>
<td>222</td>
<td>56</td>
</tr>
<tr>
<td>Mild</td>
<td>149</td>
<td>48</td>
</tr>
<tr>
<td>Total</td>
<td>449</td>
<td>61</td>
</tr>
<tr>
<td>Controls</td>
<td>179</td>
<td>45</td>
</tr>
</tbody>
</table>

*NEJM* 1981; 3: 10
Survival of Compliant versus Noncompliant HIV Infected Patients Administered Dinitrochlorobenzene (DNCB)

<table>
<thead>
<tr>
<th>Subjects</th>
<th>No. of Patients</th>
<th>Progression to AIDS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compliant</td>
<td>13</td>
<td>2 (15)</td>
</tr>
<tr>
<td>Noncompliant</td>
<td>11</td>
<td>5 (45)</td>
</tr>
</tbody>
</table>

“Patients who discontinued DNCB application appeared to have a higher rate of progression of AIDS”

Stricker et al. (1994)
CRITIQUE:

“The fact that compliant patients who follow instructions may have better whether their treatment works or not has been verified in several studies’

Bigby et al. (1996)
"Among women with a history of subfertility, prenatal use of electric blankets was associated with a more than four-fold increase in risk. Despite small numbers and the potential for recall bias, our study indicates that identifying a susceptible population may be required for detecting adverse reproductive effects of electromagnetic fields".

Li et al. (1995)
"With this sentence and in their Discussion, the authors signal that they would like readers to take this result seriously. Since electromagnetic fields are the exposure under study, attention is bound to be paid. The question is, how much attention is deserved?“

- Overall association is non-significant.
- Definition of compromised reproductive function questionable (history of unprotected intercourse for more than 12 months without becoming pregnant).
- Sparse data (only five exposed cases).
- No test of interaction performed.

Hatch (1995)